Automatic ICD Code Assignment to Medical Text with Semantic Relational Tuples

Sanqiang Zhao[1], Daqing He[1], Danchen Zhang[1], Lei Li[2], Rui Meng[1]
[1]University of Pittsburgh [2]Nanjing University of Science and Technology

Abstract
Mining the Electronic Medical Record (EMR henceforth) is growing in popularity but still lacks good methods for better understanding the text in EMR. One important task is assigning proper International Classification of Diseases (ICD henceforth, which is the code schema for EMR) code based on the narrative text of EMR document. For the task, we propose an automatic feature extraction method by means of capturing semantic relational tuples. We proved the semantic relational tuple is able to capture information at semantic level and it contribute to ICD-9 classification task in two aspects, negation identification and feature generation.

Keywords: ICD-9 Classification; text mining; EHR mining
DOI: Citation info is to be added.

Contact: saz31@pitt.edu, dah44@pitt.edu, daz45@pitt.edu, lileiwelldone@gmail.com, rum20@pitt.edu

1 Introduction

Electronic Medical Record (EMR) has become widely available in the medical domain, but mining such information with state-of-the-art natural language processing and machine-learning algorithms is still lacking. One important task is assigning proper International Classification of Diseases (ICD) code based on the narrative text in EMR.

ICD code is a standard classification schema and provides standard diagnoses and procedural tools for the collecting and reporting health information. Specifically, this paper focuses on the ninth revision of ICD code, clinical modification (ICD-9-CM henceforth).

Generally, ICD-9-CM is manually decided by human based on the narrative text of the EMR document. Concretely, nurses interview patients and write down complaints. Then a registration clerk or a trained human coder assigns the ICD-9-CM code(Tsui, Wagner, Dato, & Chang, 2001) by viewing the entire narrative text of the EMR document. However, due to the large number of patients and complaints, manually assignment of the ICD-9-CM code is a time-consuming and error-prone task. Therefore, automatic assignment can help in two aspects: (1) reducing error by supporting decision-making for human coders, and (2) speeding up code assignment and reducing cost(Ira Goldstein & Anna Arzumtsyan, 2007).

The goal of our study is to automatically assign an ICD-9-CM code to the given EMR document. We adopted a standard supervised machine-learning approach, which extracts a bunch of features from the narrative text and learns the weight of each feature from the training data set. Previous studies have examined the performance on several simple feature-extraction methods, such as bag-of-word and rule-based features. However, as the medical text in EMR is relatively simple and straightforward, it is possible to better understand the text by applying a state-of-the-art natural language technique.

Specifically, we propose an automatic feature extraction method for capturing semantic relational tuples, like <rib,pain> and <cough, persistent>. This kind of tuples are widely distributed in narrative text of EMR documents and can be used to identify their ICD-9-CM codes effectively. Compared to features in prior studies, features from our approach delves into the semantic level of medical text.

2 Approach for Semantic Tuple Extraction

Though previous studies (Crammer, Dredze, Ganchev, Talukdar, & Carroll, 2007; Farkas & Szarvas, 2008; Ira Goldstein & Anna Arzumtsyan, 2007) have demonstrated the effectiveness of different features for the ICD-9-CM classification task, most of these features require predefined extracting rules and also unable to

discover enough semantic features. Instead, we need a method for extracting more meaningful features and reducing manual effort as much as possible. Kavuluru et al.(Kavuluru, Rios, & Lu, 2015; Kavuluru, Han, & Harris, 2013) developed an automatic approach, in which they applied the part-of-speech tagger and extracted word sequence following certain part-of-speech tag patterns, such as adjective + noun, noun + noun, etc., However, the part-of-speech tagger cannot extract syntactic relationships and cannot handle modification relationships from texts.

Inspired by the idea of extracting semantic relational tuples for the product review(Qiu, Liu, Bu, & Chen, 2011), we extract the <aspect, status> tuples, in which an aspect can be a symptom, procedure, etc., and a status can be the modifier for this aspect, such as <cough, recurrent> (cough being one type of symptom while recurrent being status), <rib, pain> (rib being body part while pain being status).

Here, we discuss preliminary reasons why our method is supposed to be useful.

- Narrative text in EMR documents is simple and straightforward, few implicit opinions or rhetoric in it. Therefore, the <aspect, status> tuple is reachable and effective expression for understanding the narrative text.
- Narrative text in EMR documents is formal and the majority of expressions is standard.

3 Experiment

3.1 Experiment Setup

We evaluated effectiveness of our semantic relational tuples through classification tasks. We utilize Logistic Regression as our classifier on the CMC data set, which contains 978 radiological reports. Due to data set is skewed (data sizes for different codes are imbalanced), previous researchers and us adopt the micro-averaged f1-score as the metric.

The 10-fold cross-validation was employed in our experiment. Also, in order to reduce the random effect, we run the 10-fold cross- validation multiple times (100 times in our case) and provide the mean of micro-averaged f1-scores.

3.2 Identify Negation

Based on prior studies, negation expression, such as "no focal pneumonia", should not be considered when assigning ICD code(Ira Goldstein & Anna Arzumtsyan, 2007). Traditional approach for identifying negation expression is through Negex algorithm, which is rule-based algorithm and incorporates many hand-built rules. However, our method automatically identify negations without predefined rules but some opinion words, such as "no". In this section, we compare classification performance by removing negation detected by different negation detection methods.

Negation Detection Method	Nothing Applied	Negex Algorithm	Our method
f1-score	0.827	0.842	0.844

Table 1: Classification Performance based on different Negation Detection Algorithm

Based on Table 1, both negation detection algorithm would detect negation for improving classification performance. Our method performs similar as Negex algorithm. However, Negex algorithm benefits from many hand-built rules and major inaccuracy comes from the lack and conflict of hand-built rules. Our method keeps away from the limitations.

3.3 Selected Features

By further proving effectiveness of our semantic feature tuple, feature selection is employed. Note that we have totally 2700 semantic features generated by our approach and we rank them by inverse document frequency (IDF) (Yang & Pedersen, 1997) and add them into model gradually.

Number of Features	500	1000	1500	2000	2500	2700
f1-score	0.842	0.844	0.842	0.844	0.845	0.854

Table 2: Classification Performance by Adding Feature Gradually

Table 2 shows trend of the performance by adding 500, 1000, 1500, 2000, 2500 and 2700 semantic features respectively. Initially the improvement on classification performance is little but last 700 features contributes to 1 percentage of improvement.

4 Conclusion

Our feature extraction approach for capturing semantic relational tuple significantly improves performance of ICD-9-CM code assignment task. The most important reason is our approach delve into semantic level of text. The semantic relational tuple is widely distributed in narrative text of EMR documents and can be used to identify the ICD-9-CM codes effectively. Experiment result shows that our approach could achieve significant improvement on f1-score by more than 1 percentage. Another contribution is our method can also identify negation expression without hand-built rules.

References

Crammer, K., Dredze, M., Ganchev, K., Talukdar, P. P., & Carroll, S. (2007). Automatic code assignment to medical text. In *Proceedings of the workshop on bionlp 2007: Biological, translational, and clinical language processing* (pp. 129–136).

Farkas, R., & Szarvas, G. (2008). Automatic construction of rule-based icd-9-cm coding systems. *BMC bioinformatics*, *9*(Suppl 3), S10.

Ira Goldstein, M., & Anna Arzumtsyan, M. (2007). Three approaches to automatic assignment of icd-9-cm codes to radiology reports.

Kavuluru, R., Han, S., & Harris, D. (2013). Unsupervised extraction of diagnosis codes from emrs using knowledge-based and extractive text summarization techniques. In *Canadian conference on artificial intelligence* (pp. 77–88).

Kavuluru, R., Rios, A., & Lu, Y. (2015). An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, *65*(2), 155–166.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expansion and target extraction through double propagation. *Computational linguistics*, *37*(1), 9–27.

Tsui, F.-C., Wagner, M. M., Dato, V., & Chang, C. (2001). Value of icd-9 coded chief complaints for detection of epidemics. In *Proceedings of the amia symposium* (p. 711).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412–420).